Inferring the selective pressures acting on insertions and deletions in the great tit genome

Henry Barton

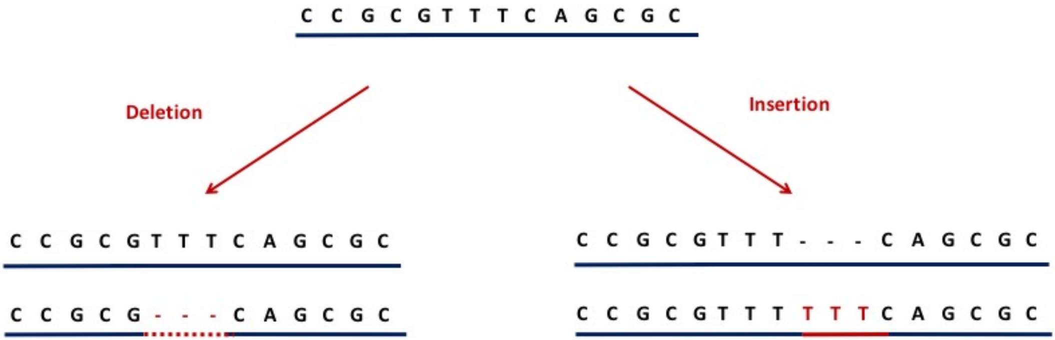
The University of Sheffield

22/08/18

Introduction

Insertions and deletions

- short INDELs: sections of DNA < 50bp that are deleted or inserted in a genome


C C G C G T T T C A G C G C
Deletion
Insertion
C C G C G T T T C A G C G C
C C G C G - - - C A G C G C
C C G C G T T T - - - C A G C G C
C C G C G T T T T T T C A G C G C

INDELs often overlooked

- Disproportionately occur in repetitive sequence
- Hard to align
- Often occur in hotspots
- 1/8 as frequent as SNPs in humans

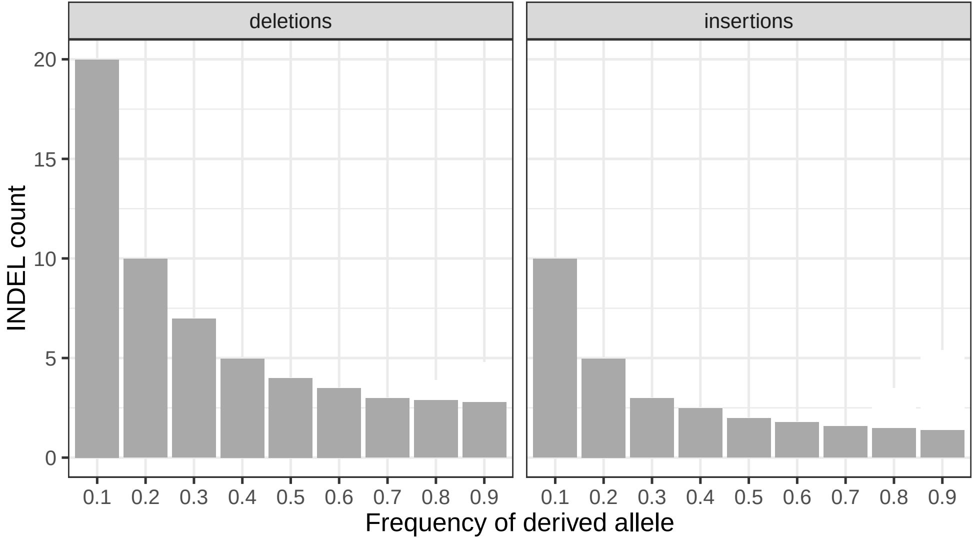
(Earl et al., 2014; Montgomery et al., 2013)

The importance of INDELs in genome evolution

- Contribute more to sequence divergence, in terms of the number of base differences, than SNPs
- Influence genome size:
  - low deletion rate $\rightarrow$ large genomes?
  - high deletion rate $\rightarrow$ compact genomes?
- Selection on insertions to maintain minimum intron size?
- Picture is complicated by errors identifying ancestral states

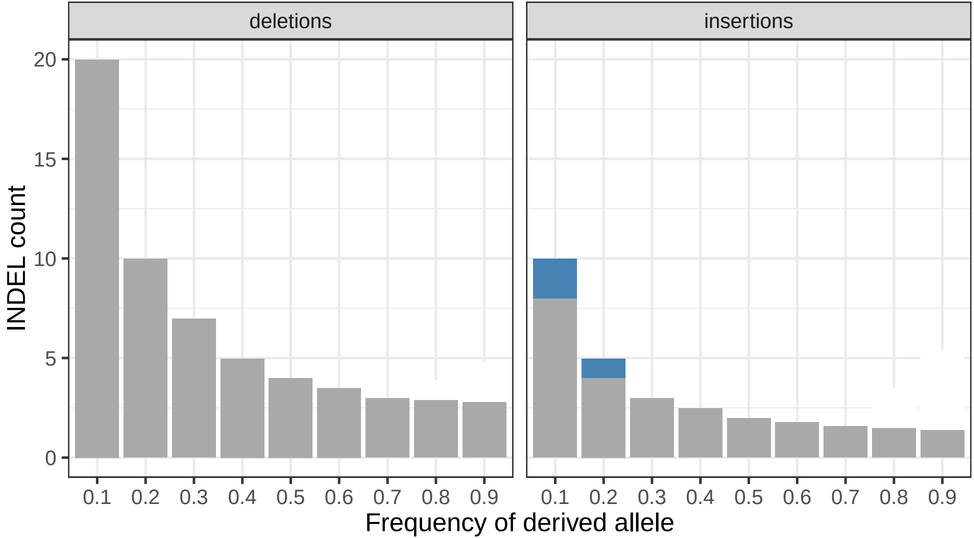
(Britten, 2002; Leushkin and Bazykin, 2013; Nam and Ellegren, 2012; Ometto et al., 2005; Sun et al., 2012)

Importance of polarisation error



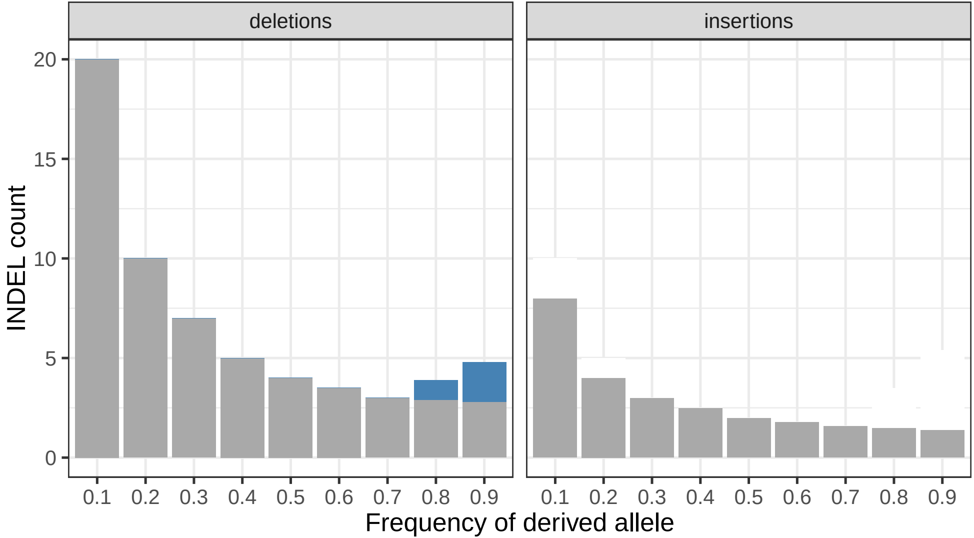
(see Hernandez et al., 2007)

Importance of polarisation error



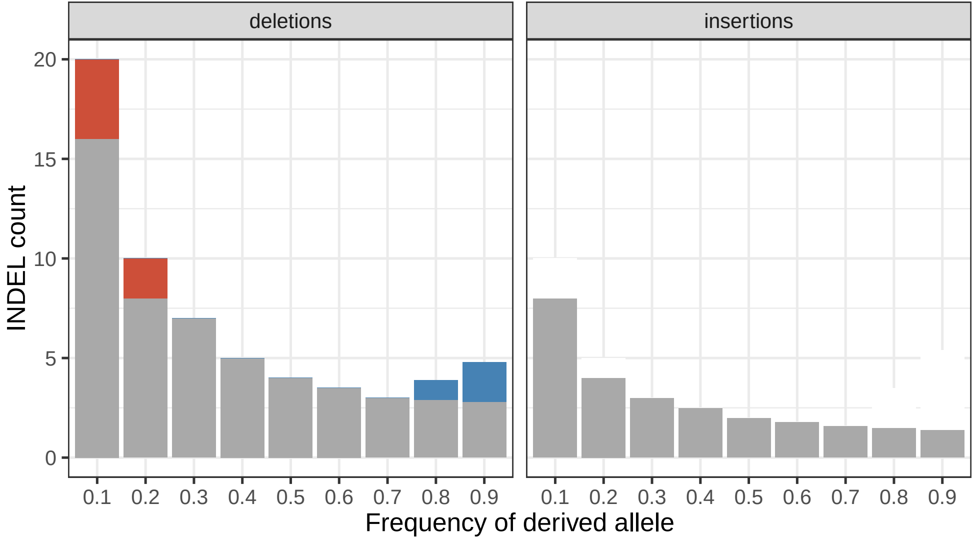
(see Hernandez et al., 2007)

Importance of polarisation error



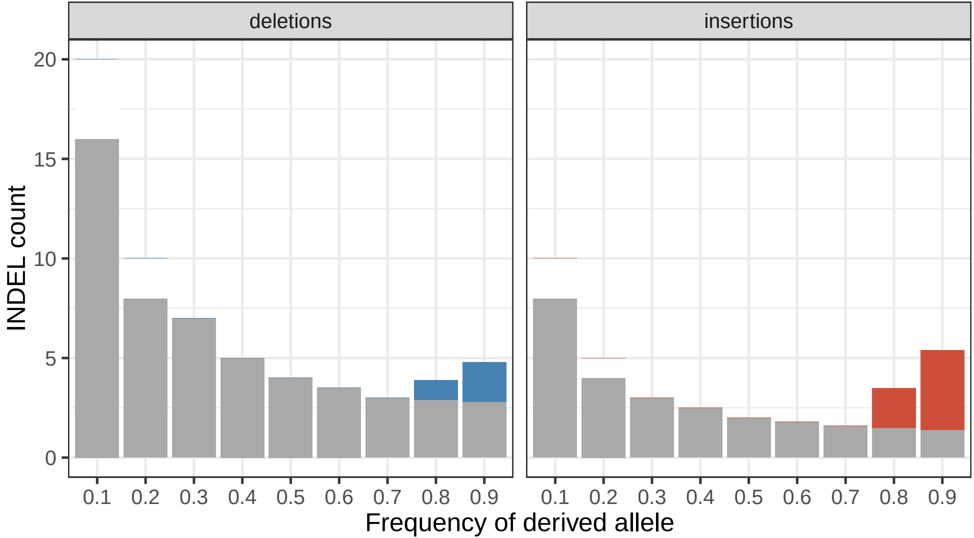
(see Hernandez et al., 2007)

Importance of polarisation error



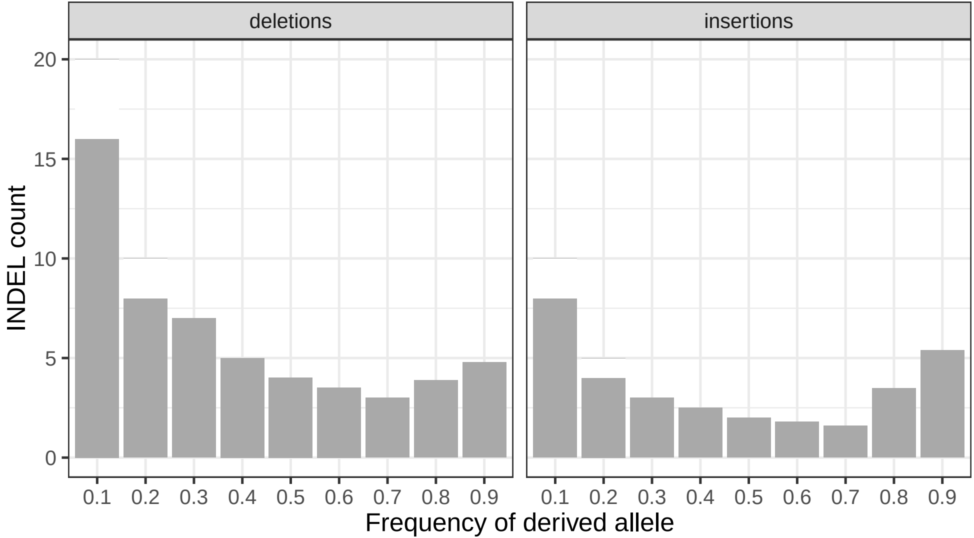
(see Hernandez et al., 2007)

Importance of polarisation error



(see Hernandez et al., 2007)

Importance of polarisation error



(see Hernandez et al., 2007)

Aims

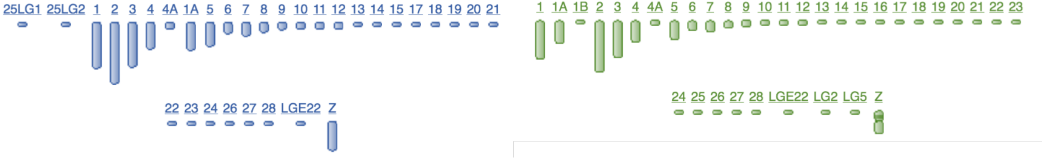
Overcome confounding affect of polarisation error

Quantify how natural selection shapes INDEL diversity in the great tit (*Parus major*)

1. within coding regions
2. in non-coding regions

Advantages of an avian system

- Conserved karyotype and synteny - good for alignments
- Genomes consist of few large macrochromosomes and many small microchromosomes
- Results in a highly dynamic recombination landscape - power to associations with recombination

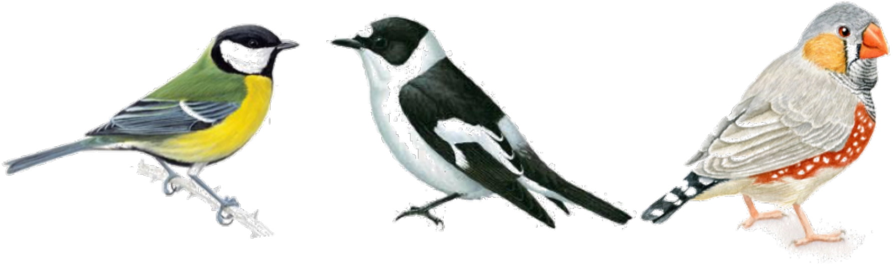

25LG1 25LG2 1 2 3 4 4A 1A 5 6 7 8 9 10 11 12 13 14 15 17 18 19 20 21
22 23 24 26 27 28 LGE22 Z
1 1A 1B 2 3 4 4A 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24 25 26 27 28 LGE22 LG2 LG5 Z

(van Oers et al., 2014; Stapley et al., 2008)

Data

Sample and pipeline

- 10 european great tit males (Corcoran et al., 2017)
- high coverage (44x)
- variant calling with GATK
- multispecies alignment between zebra finch, flycatcher and great tit
- parsimony based polarisation



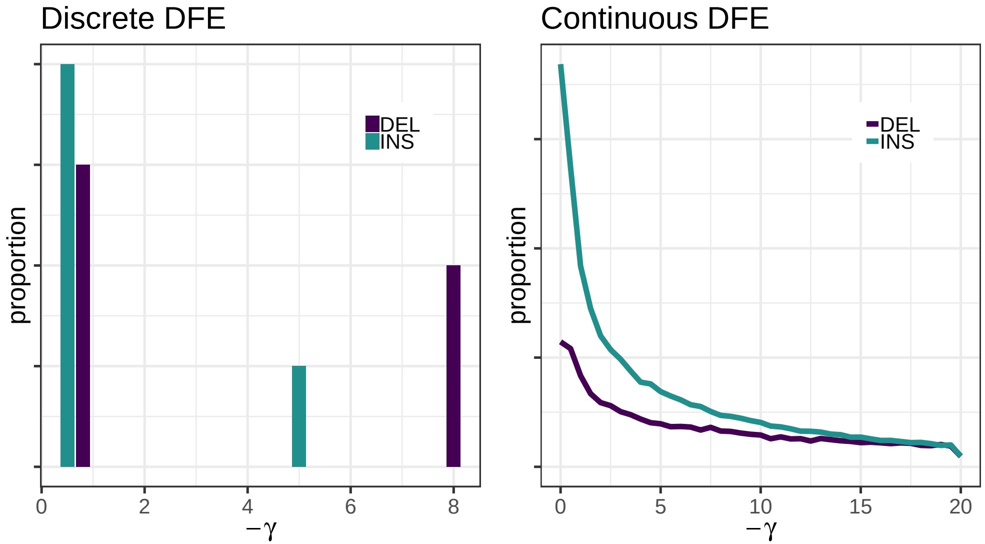
The model - ‘anavar’

A novel maximum likelihood approach

- takes the unfolded site frequency spectrum
- estimates for both insertions and deletions:
  - mutation rate ($\theta = 4N_e\mu$)
  - the distribution of fitness effects (DFE)
  - polarisation error
- Controls for demography using neutral sites (Eyre-Walker et al., 2006)
- Applicable to both INDELs and SNPs or a combination

(Barton and Zeng, MBE, 2018)

the model DFEs

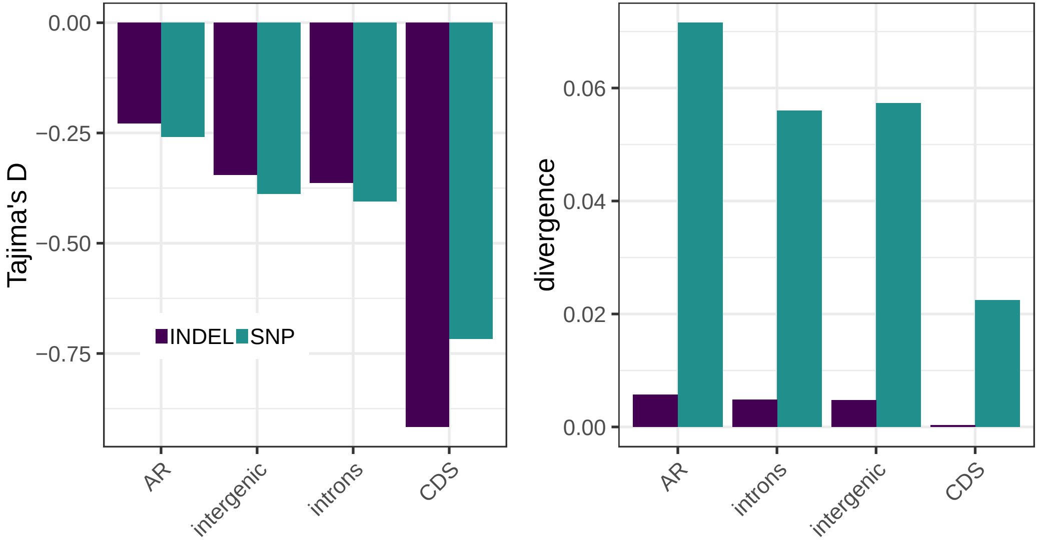

Discrete DFE

Continuous DFE

(Barton and Zeng, MBE, 2018)

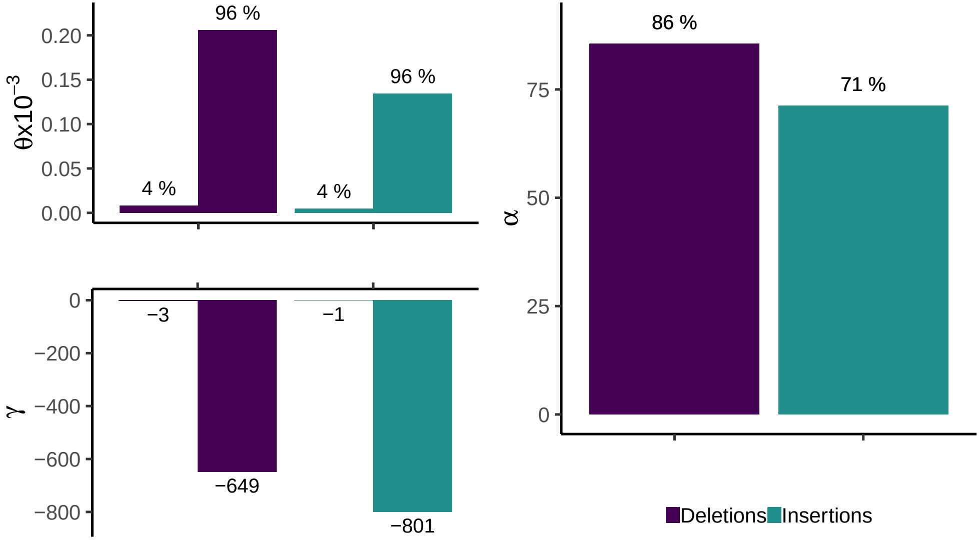
Dataset summary

Regional variation in purifying selection



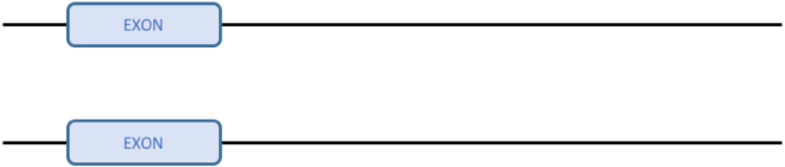
Coding INDELs

Polymorphic INDELs predominantly strongly deleterious

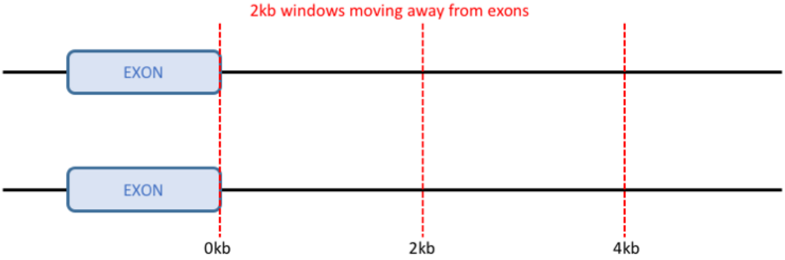


Moving away from coding regions

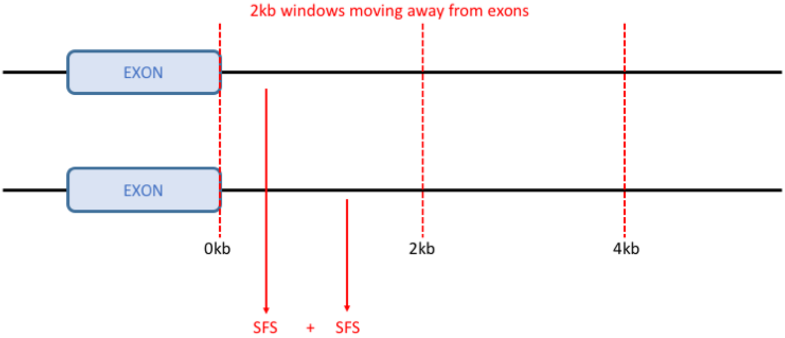
Approach



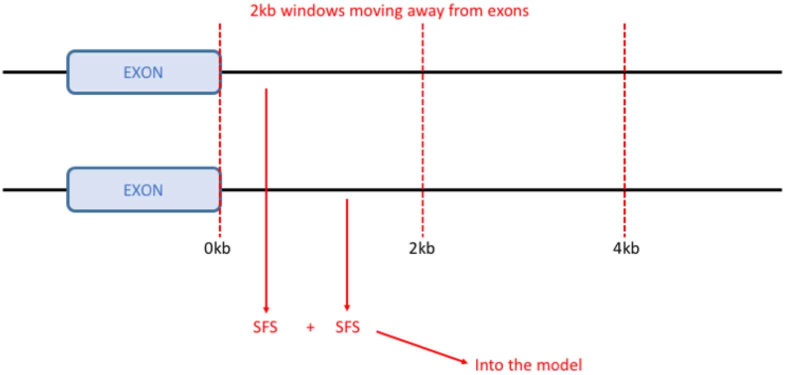
Approach



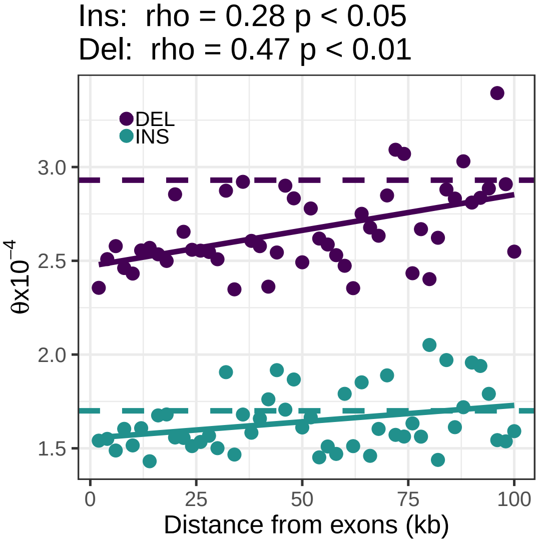
Approach


2kb windows moving away from exons
EXON
EXON
0kb
2kb
4kb
SFS + SFS

Approach

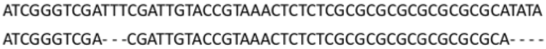

2kb windows moving away from exons
EXON
EXON
0kb
2kb
4kb
SFS
+
SFS
Into the model

Evidence for linked selection



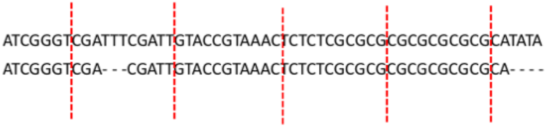
Recombination Analyses

Getting the data



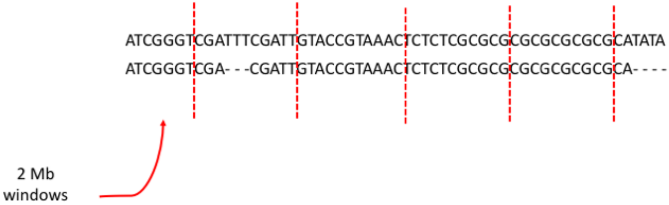
```
ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCATATA
ATCGGGTCGA- - -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCA- - - -
```

Getting the data



```
ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCATATA
ATCGGGTCGA- - -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCA- - - -
```

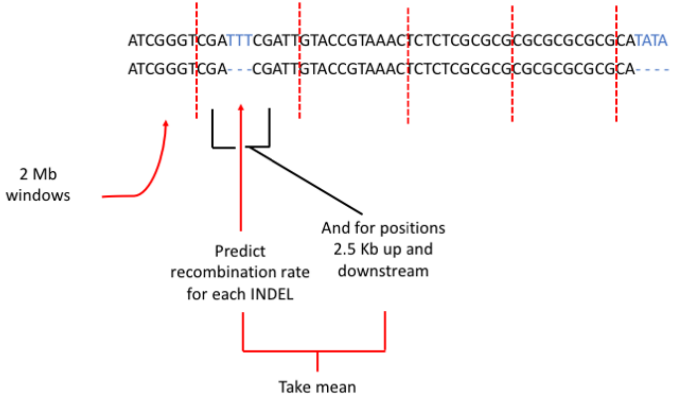
Getting the data



```
ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCATATA
ATCGGGTCGA- - -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCA- - - -
```

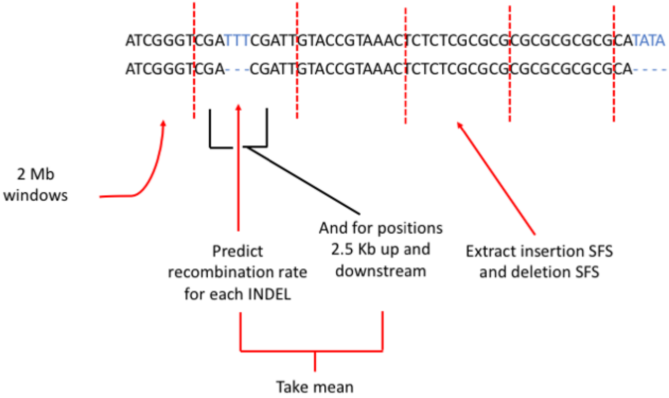
2 Mb
windows

Getting the data



ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCATATA
ATCGGGTCGA- - -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCA- - - -
2 Mb windows
Predict recombination rate for each INDEL
And for positions 2.5 Kb up and downstream
Take mean

Getting the data



ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCGCATATA

ATCGGGTCGA- - -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCGCA- - - -

2 Mb windows

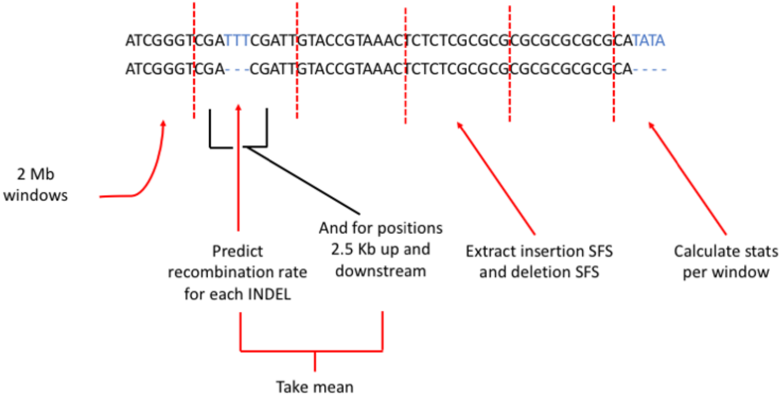
Predict recombination rate for each INDEL

And for positions 2.5 Kb up and downstream

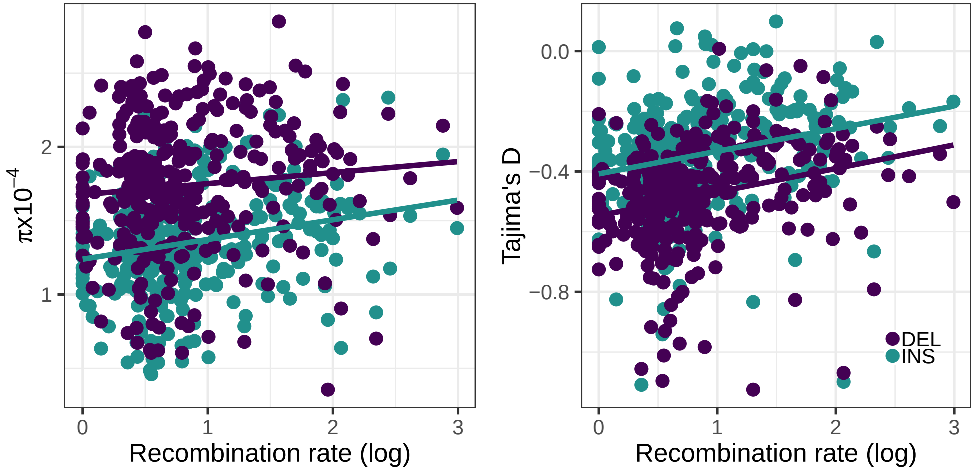
Extract insertion SFS and deletion SFS

Take mean

Getting the data


ATCGGGTCGATTTCGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCATATA
ATCGGGTCGA- - -CGATTGTACCGTAAACTCTCTCGCGCGCGCGCGCGCA- - - -
2 Mb windows
Predict recombination rate for each INDEL
And for positions 2.5 Kb up and downstream
Extract insertion SFS and deletion SFS
Calculate stats per window
Take mean

Association between diversity and recombination



Ins: rho = 0.18 p < 0.05
Del: rho = 0.12 p < 0.01

Ins: rho = 0.3 p < 0.01
Del: rho = 0.33 p < 0.01

Round up

Conclusion

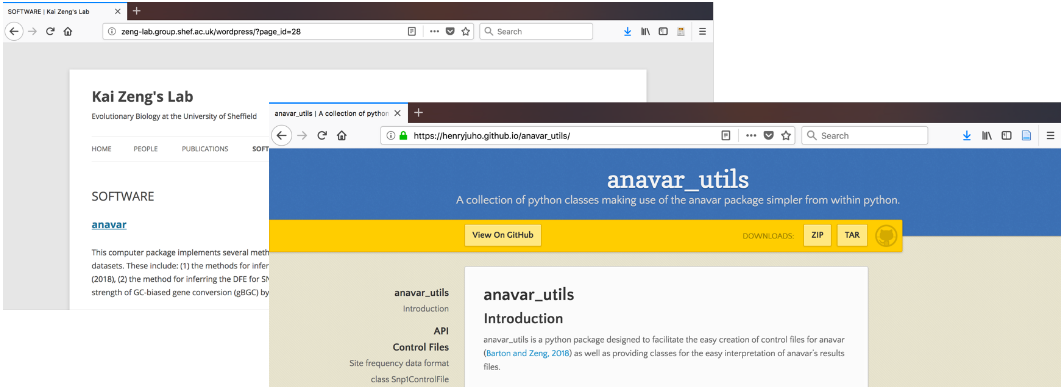
- INDELs in genes mostly extremely deleterious - 96%
- Remainder are weakly deleterious - deletions more so
- α estimate at 71% and 86% for insertions and deletions
- Regions adjacent to exons, and areas of low recombination, have reduced INDEL diversity - genetic hitch-hiking
- Extends over relatively large distance 0-100kb

Next steps

- Interesting to investigate if reduced diversity is due to positive selection or purifying selection
- Look at whether efficacy of selection on INDELs is higher in regions with higher N_e

Plug for the model

- User friendly computer package - anavar - `http://zeng-lab.group.shef.ac.uk`
- Methods are applicable to both INDELs and SNPs or a combination
- Code for integration with python - `https://henryjuho.github.io/anavar_utils/`



SOFTWARE | Kai Zeng's Lab
zeng-lab.group.shef.ac.uk/wordpress/?page_id=28
Search
HOME
PEOPLE
PUBLICATIONS
SOFT

Kai Zeng's Lab

Evolutionary Biology at the University of Sheffield

SOFTWARE

anavar

This computer package implements several meth
datasets. These include: (1) the methods for infer
(2018), (2) the method for inferring the DFE for SN
strength of GC-biased gene conversion (gBGC) by

anavar_utils | A collection of python
https://henryjuho.github.io/anavar_utils/
Search
View On GitHub
DOWNLOADS:
ZIP
TAR
anavar_utils
Introduction
API
Control Files
Site frequency data format
class Snp1ControlFile

anavar_utils

A collection of python classes making use of the anavar package simpler from within python.

anavar_utils

Introduction

anavar_utils is a python package designed to facilitate the easy creation of control files for anavar (Barton and Zeng, 2018) as well as providing classes for the easy interpretation of anavar's results files.

Questions?